

Ĥ: DISK ARRAY DATA LAYOUT TOLERATING MULTIPLE FAILURES

Barbara Theodorides & Walt Burkhard

Computer Science and Engineering Department
University of California, San Diego



- Archival storage reliability requirements.
- \hat{B} data layout ... perfect one – factorizations.
- Performance via simulation.
- Conclusions and future work.

Storage system reliability I



■ Reliability parameters

disk failure rate $\lambda \approx 1 / 10^6$ hour $< 1 / 100$ years.

disk repair rate $\mu \approx 1 /$ hour.

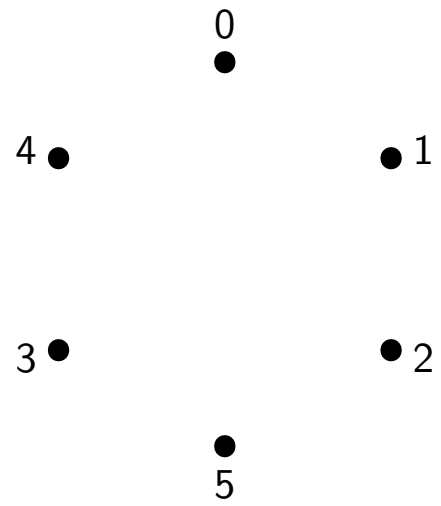
	disks	\approx MTTDL	disk capacity
mirroring	2	$\frac{1}{2\lambda} \left(\frac{\mu}{\lambda} \right)$	1
RAID 5	$n \geq 2$	$\frac{1}{n(n-1)\lambda} \left(\frac{\mu}{\lambda} \right)$	$n - 1$
RAID 6	$n \geq 3$	$\frac{1}{n(n-1)(n-2)\lambda} \left(\frac{\mu}{\lambda} \right)^2$	$n - 2$
$2n$ -mirroring	$2n \geq 2$ (even)	$\frac{1}{(2n)!\lambda} \left(\frac{\mu}{\lambda} \right)^{2n-1}$	1
\hat{B}	$2n \geq 4$ (even)	$\frac{1}{(2n)!\lambda} \left(\frac{\mu}{\lambda} \right)^{2n-2}$	2



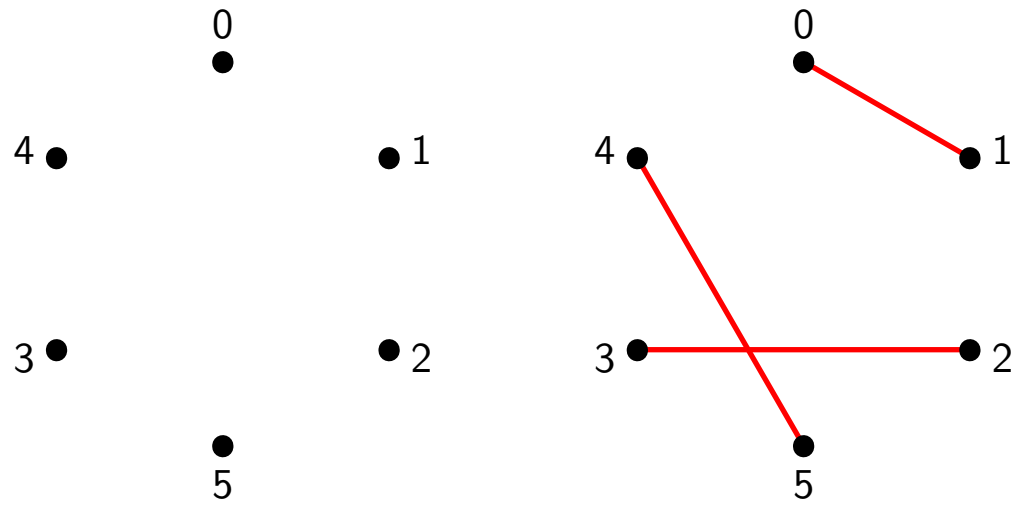
- \hat{B} with $2n$ disks is based upon a perfect one-factorization of a complete graph on $2n+2$ nodes.

A **factor** of graph G is a spanning subgraph and a **one-factor** of G is a one-regular spanning subgraph. A **factorization** is a set of factors which are pairwise edge disjoint whose union is G . A factorization consisting of one-factors is a **perfect one-factorization** if any distinct pair of factors induces a Hamiltonian cycle.

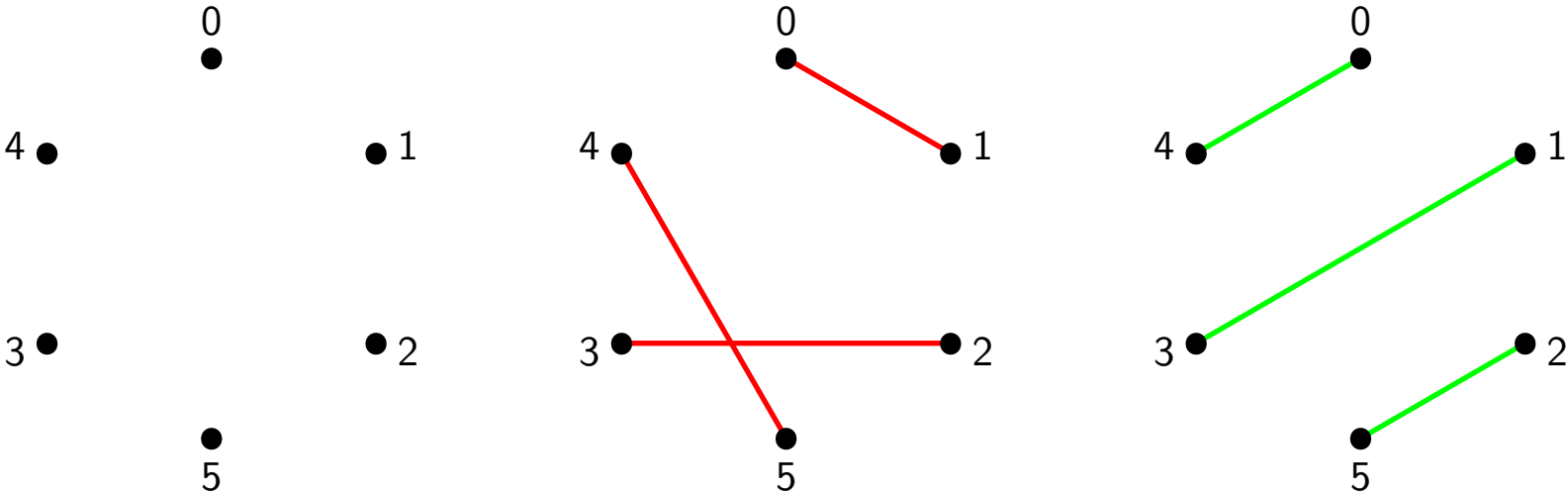
\hat{B} Data Layout II



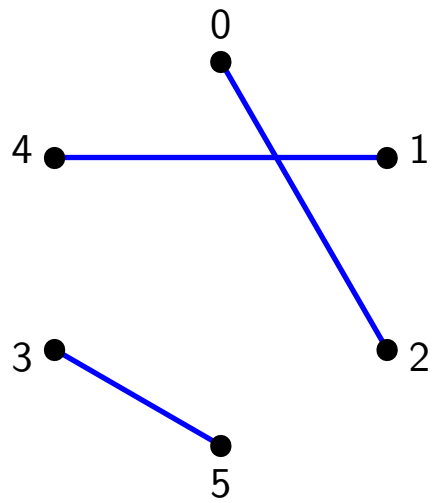
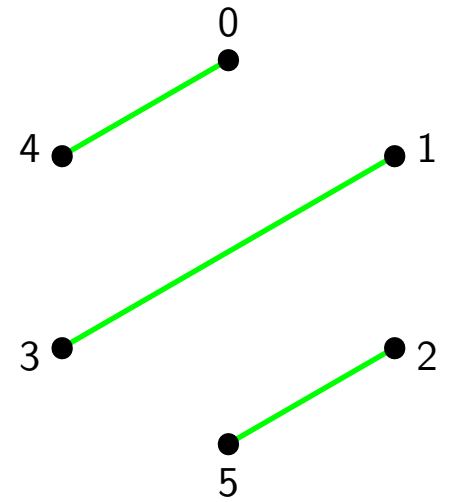
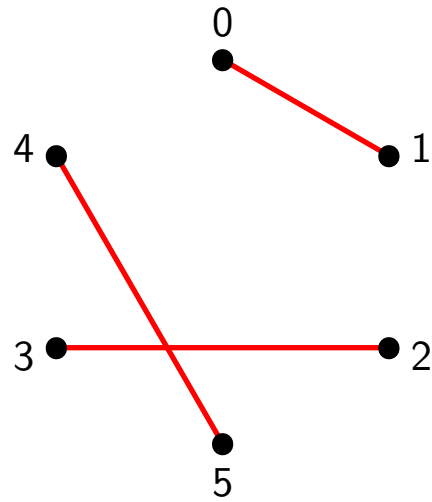
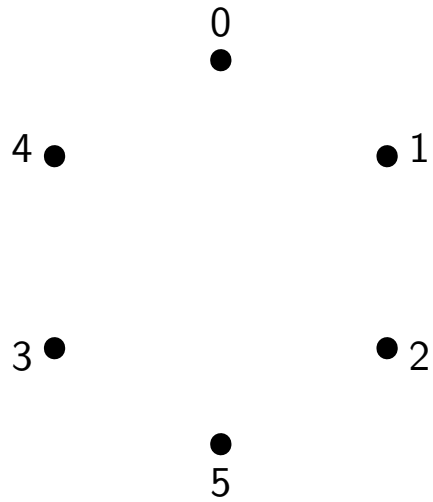
\hat{B} Data Layout III



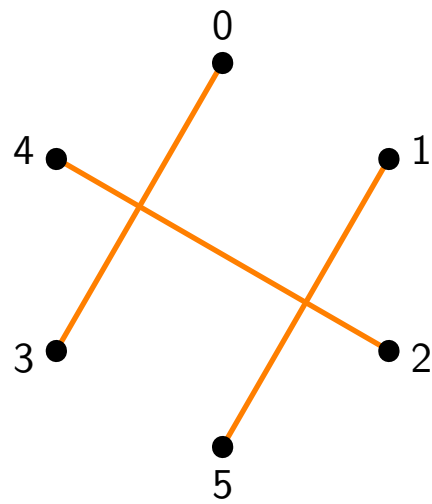
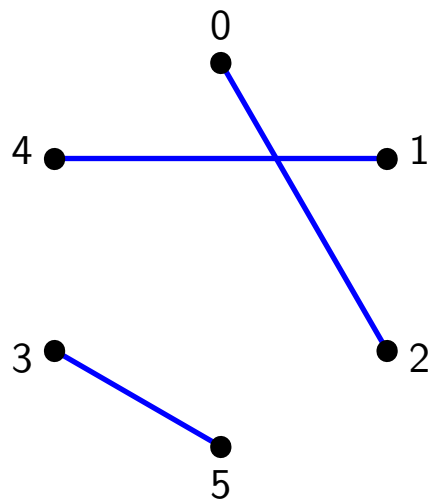
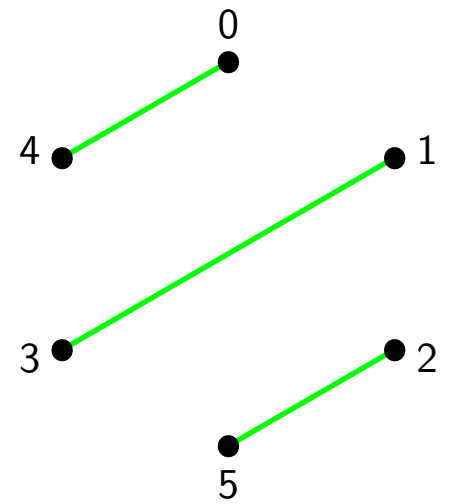
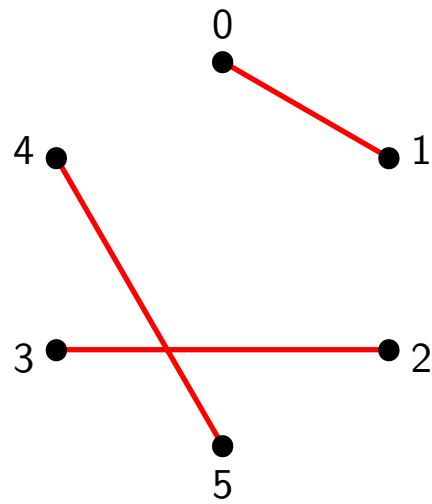
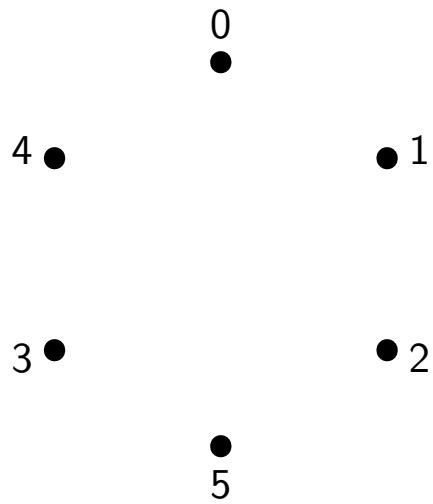
\hat{B} Data Layout IV



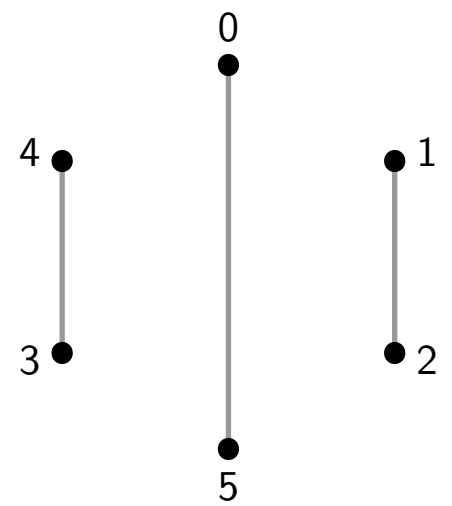
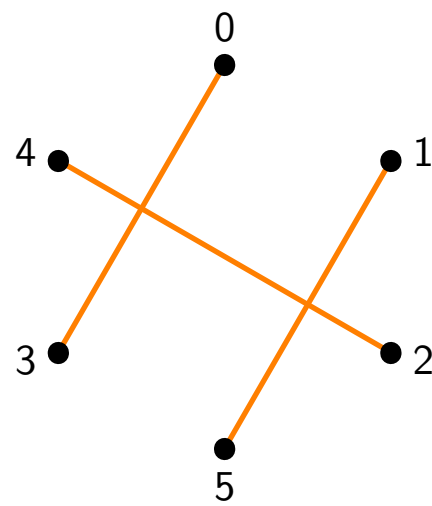
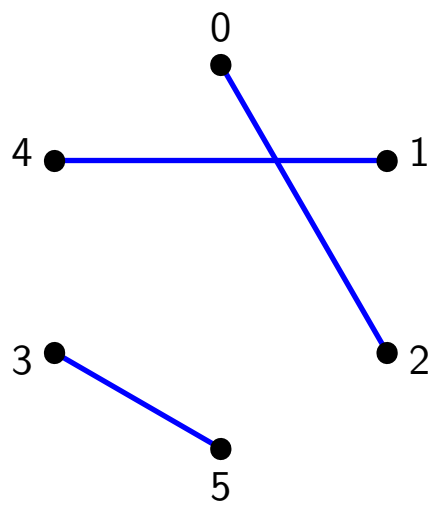
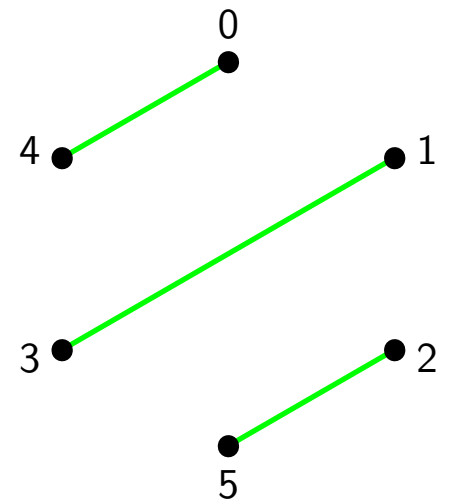
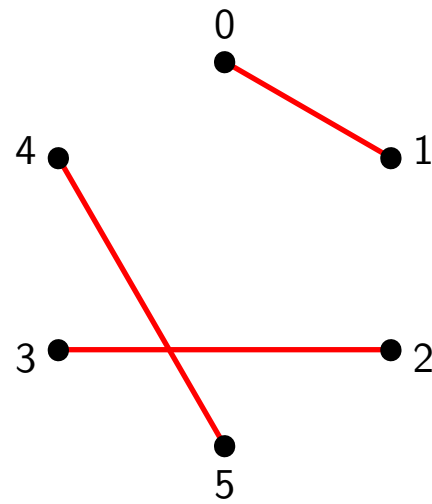
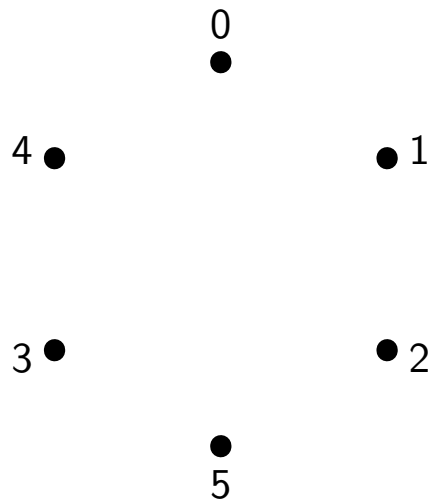
\hat{B} Data Layout V



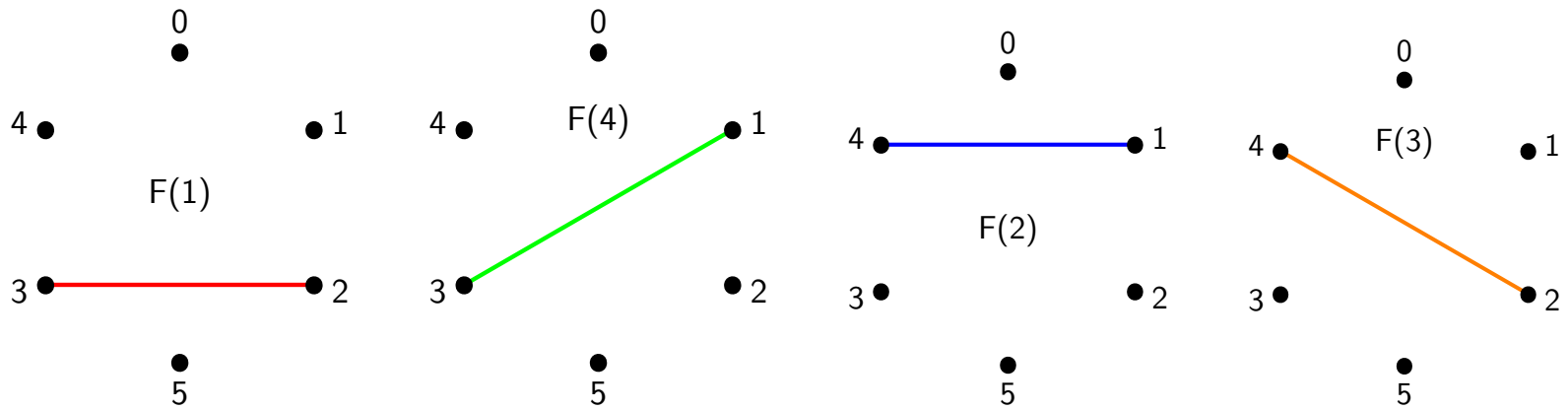
\hat{B} Data Layout VI



\hat{B} Data Layout VII



\hat{B} Data Layout VIII



disk 1	disk 2	disk 3	disk 4
d_1	d_2	d_3	d_4
$d_2 \oplus d_3$	$d_1 \oplus d_4$	$d_2 \oplus d_4$	$d_1 \oplus d_3$

\hat{B}_4 data layout



For odd-prime p , complete graphs K_{p+1} and K_{2p} possess perfect one-factorizations. The $2n = p - 1$ or $2p - 2$ one-factors are each modified by removing two edges to obtain the parity subsets.

There are a few more one-factorizations known – e.g. for 36 nodes.

\hat{B} Data Layout X



- \hat{B}_{2n} structural properties: $2n$ disks, $n - 1$ check stripe units per data stripe unit, utilizes the minimal amount of additional storage to ensure reconstruction for up to $2n - 2$ simultaneous disk failures. The number of redundant bits modified by changing a single client data bit is minimal.

disk 1	disk 2	disk 3	disk 4	disk 5	disk 6	disk 7	disk 8
d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
$d_2 \oplus d_3$	$d_1 \oplus d_6$	$d_1 \oplus d_8$	$d_1 \oplus d_3$	$d_2 \oplus d_8$	$d_1 \oplus d_5$	$d_1 \oplus d_4$	$d_1 \oplus d_7$
$d_4 \oplus d_5$	$d_4 \oplus d_8$	$d_2 \oplus d_5$	$d_2 \oplus d_7$	$d_3 \oplus d_7$	$d_2 \oplus d_4$	$d_3 \oplus d_6$	$d_2 \oplus d_6$
$d_6 \oplus d_7$	$d_5 \oplus d_7$	$d_4 \oplus d_7$	$d_6 \oplus d_8$	$d_4 \oplus d_6$	$d_3 \oplus d_8$	$d_5 \oplus d_8$	$d_3 \oplus d_5$



- Raidframe simulation enhancements include:
 - \hat{B} data layout and reconstruction schemes,
 - Provide for $2n-2$ concurrent disk failures,
 - Disk drive technology – IBM 18LZX.
- 8 disk RAID 5 and \hat{B} stripes, 7 and 9 disk EVENODD stripes.
- synthetic read and write workloads, 16 sector stripe units, 8KB, 32KB, & 96KB operations.

Performance via simulation II

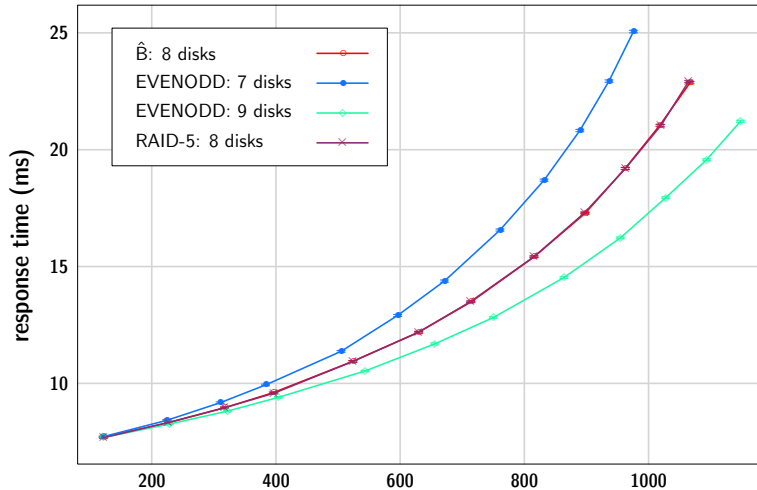


Figure 6a: 8KB read operations per second: all disks operational

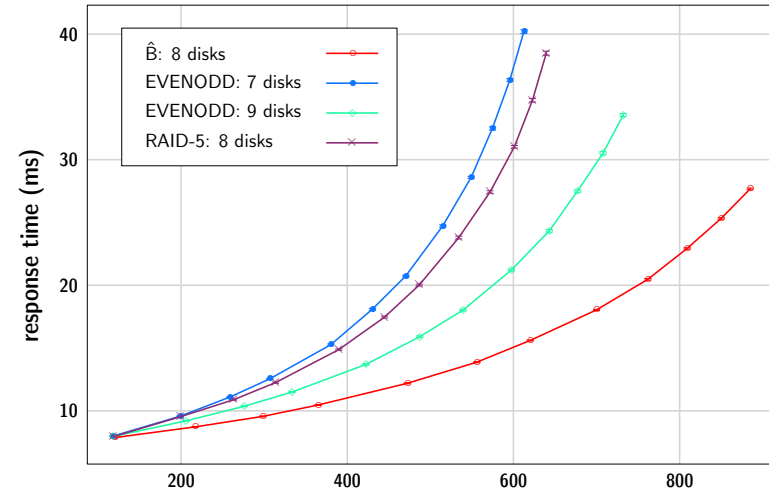


Figure 6b: 8KB read operations per second: one failed disk

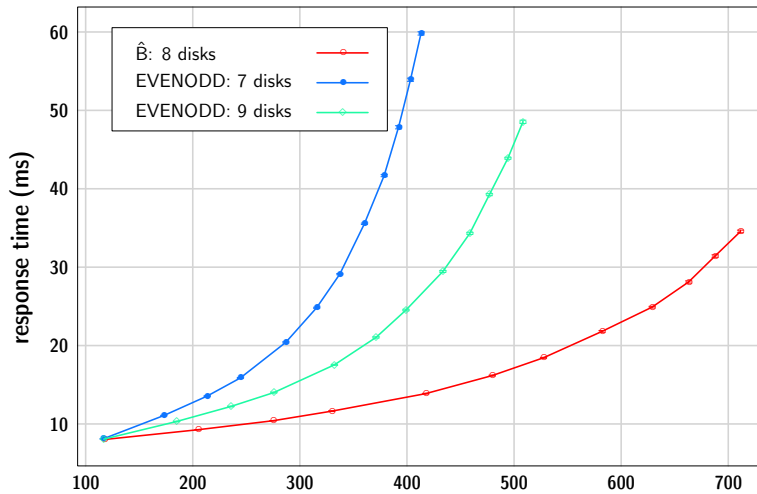


Figure 6c: 8KB read operations per second: two failed disks

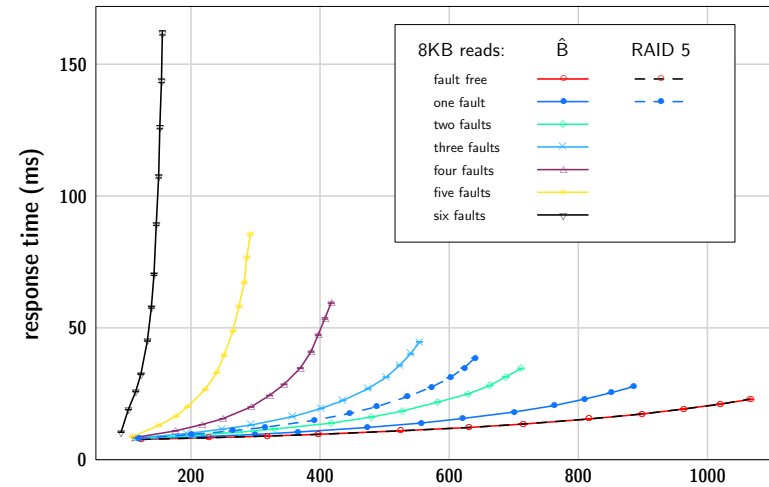


Figure 6d: 8KB read operations per second

Performance via Simulation III

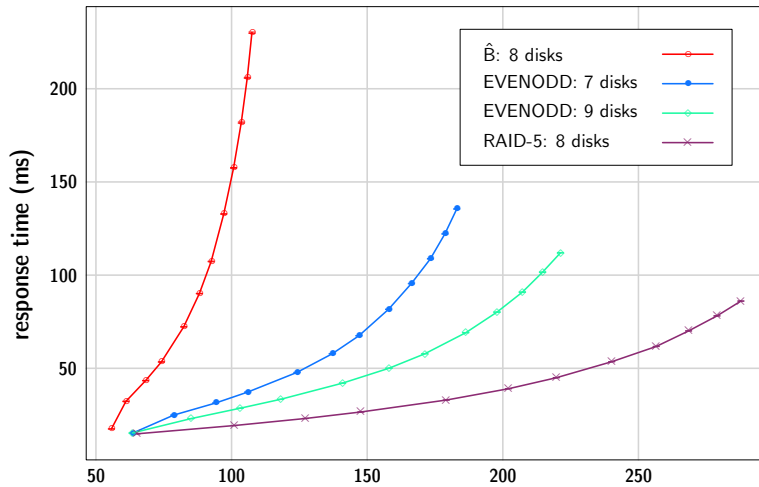


Figure 8a: 8KB write operations per second: all disks operational

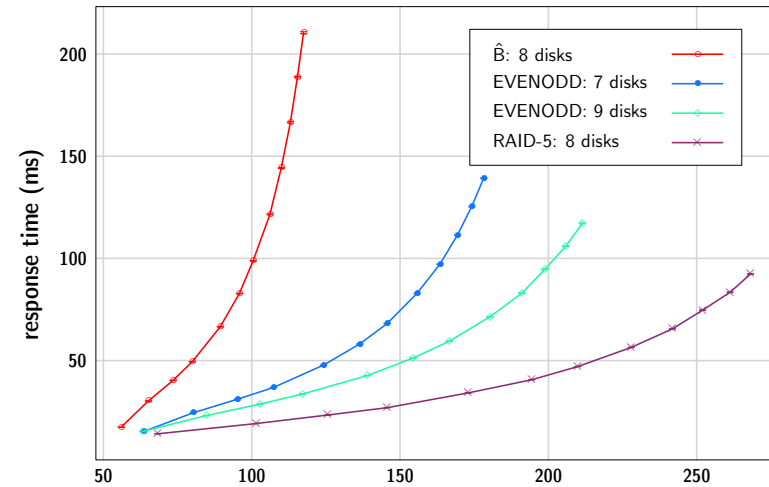


Figure 8b: 8KB write operations per second: one failed disk

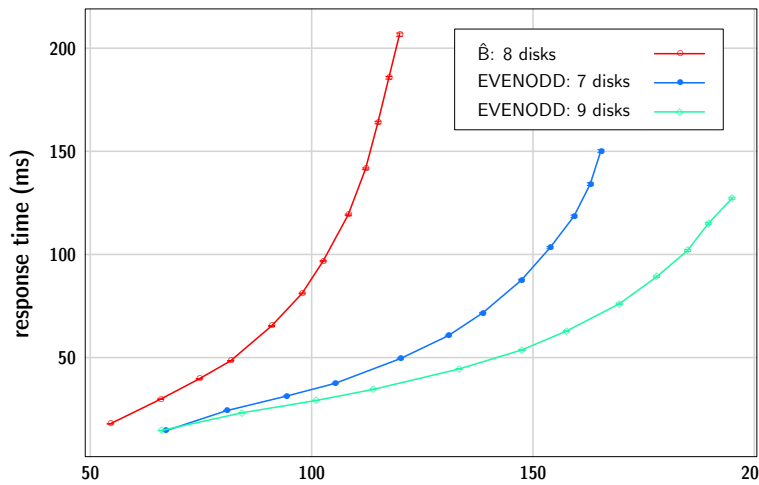


Figure 8c: 8KB write operations per second: two failed disks

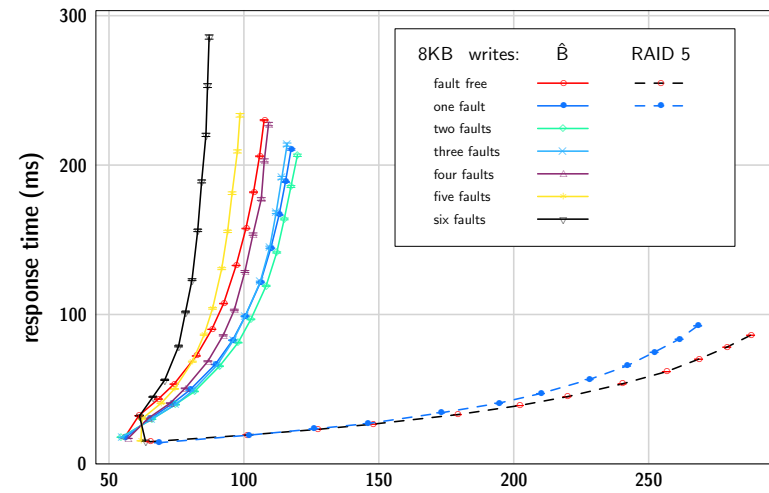


Figure 8d: 8KB write operations per second



- \hat{B} read performance
 - no failure: identical to RAID5 and EVENODD.
 - one, two failures: better than RAID5 or EVENODD.
- \hat{B} write performance
 - much worse than RAID 5 or EVENODD.
- \hat{B} is a possible data layout candidate for readonly archival storage.



- Performance comparison via simulation –

\hat{B}_{2n}

$2n$ -mirroring.

self-adjusting $2n$ -mirroring.

- Alternate data layouts possibly utilizing perfect factorizations within hypergraphs to trade storage capacity with reliability.